Car Price Prediction Report

1. Dataset Description

Dataset: CarPrice_Dirty.csv
Dataset link: Car price prediction
Features: Make, Year, Engine Size, Mileage, Fuel Type, Transmission
Target Variable: Price (continuous)
Objective: Predict car prices using technical and categorical features

The dataset was downloaded from Kaggle and was initially almost clean. To demonstrate and evaluate different data cleaning techniques, artificial issues were introduced into the dataset, including missing values (nulls), duplicate records, outliers, and garbage values.

2. Data Preparation

- Loaded dataset into a DataFrame
- Removed duplicates and irrelevant columns
- Handling Garbage Values by converting it with NaN
- Fill null values with mean , median ,mode
- Remove outliers in Price, Mileage, and Engine Size
- Converted categorical features (Make, Fuel Type, Transmission) to numeric using One-Hot Encoding

3. Exploratory Data Analysis (EDA)

- **Shape:** 991 rows × 11 columns
- **Price Mean:** ~$25,073
- **Price Std:** ~$5,023
  Observations:
  - Engine Size positively correlated with Price
  - Mileage negatively correlated with Price
  - No strong correlations among independent features no multicollinearity issues.
  - Price distribution is slightly right-skewed, with a few high-value outliers.
  - Boxplots reveal that **Fuel Type** and **Transmission** categories influence Price distributions.

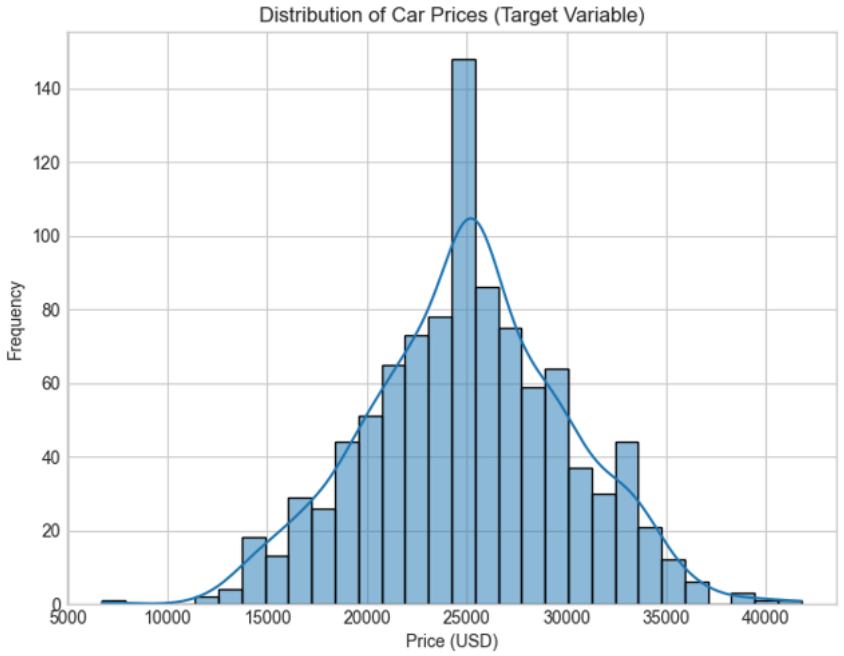
4. Visualizations

Histogram

The histogram shows the distribution of car prices after data cleaning and outlier removal

The prices are slightly right-skewed, which is typical for price-related data

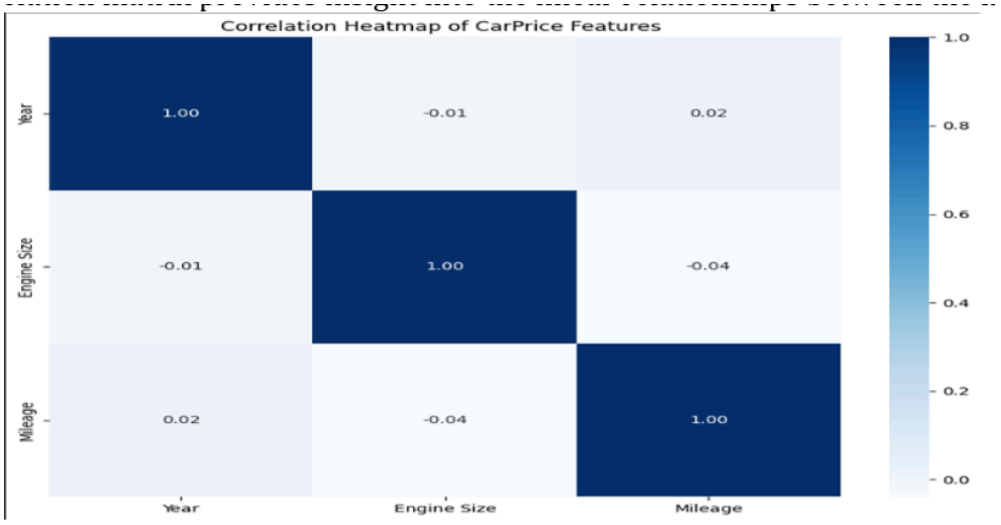
Most cars priced between ~$15,000–$30,000



Heatmap

Feature Correlation Analysis

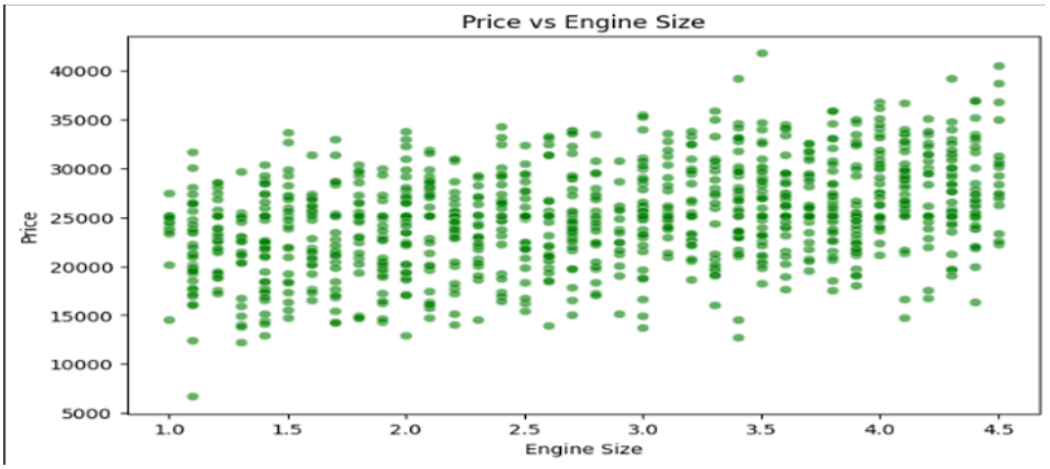
The correlation matrix provides insight into the linear relationships between the numerical features



scatter plot

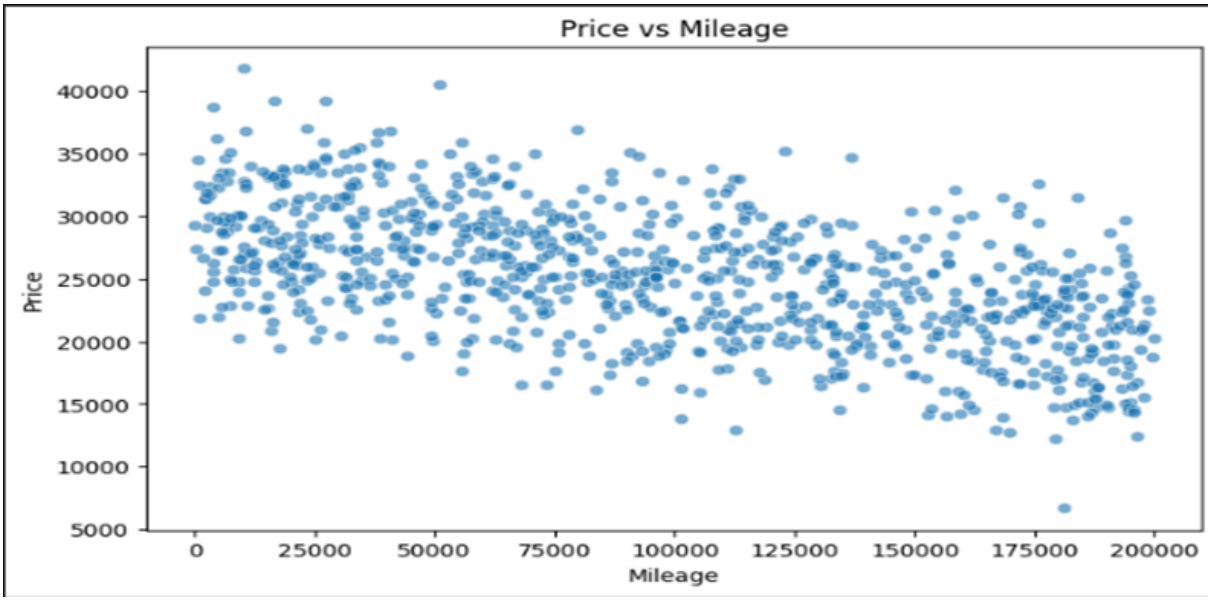
Price vs. Engine Size Visualization

visually confirms the strong positive relationship between Engine Size and Price, which is a key driver for the model's predictions.



Price vs. Mileage Visualization

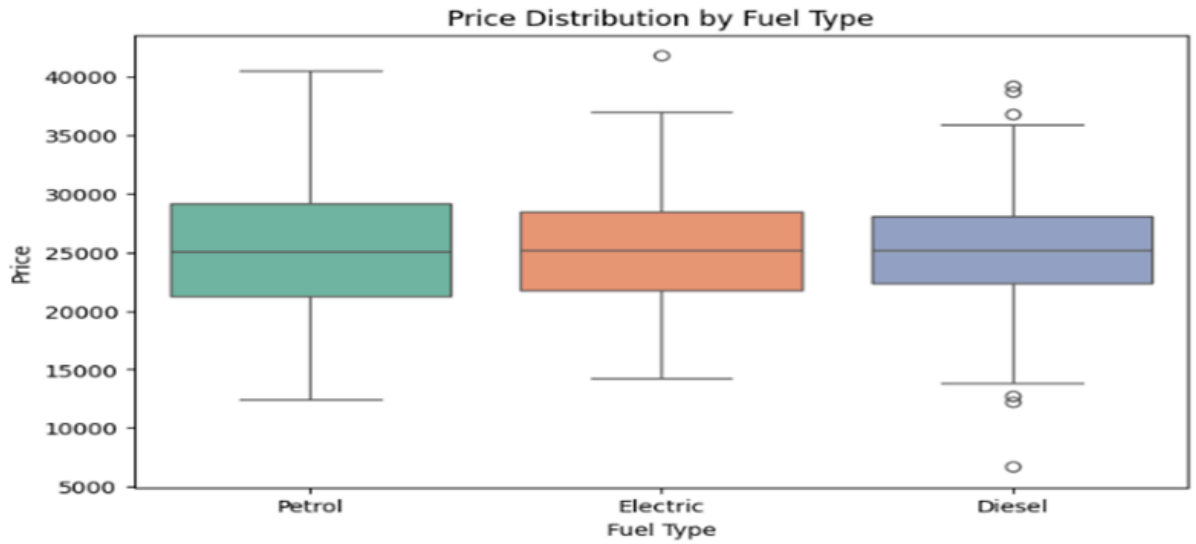
visually confirms the Negative relationship between and Mileage Price



Boxplot

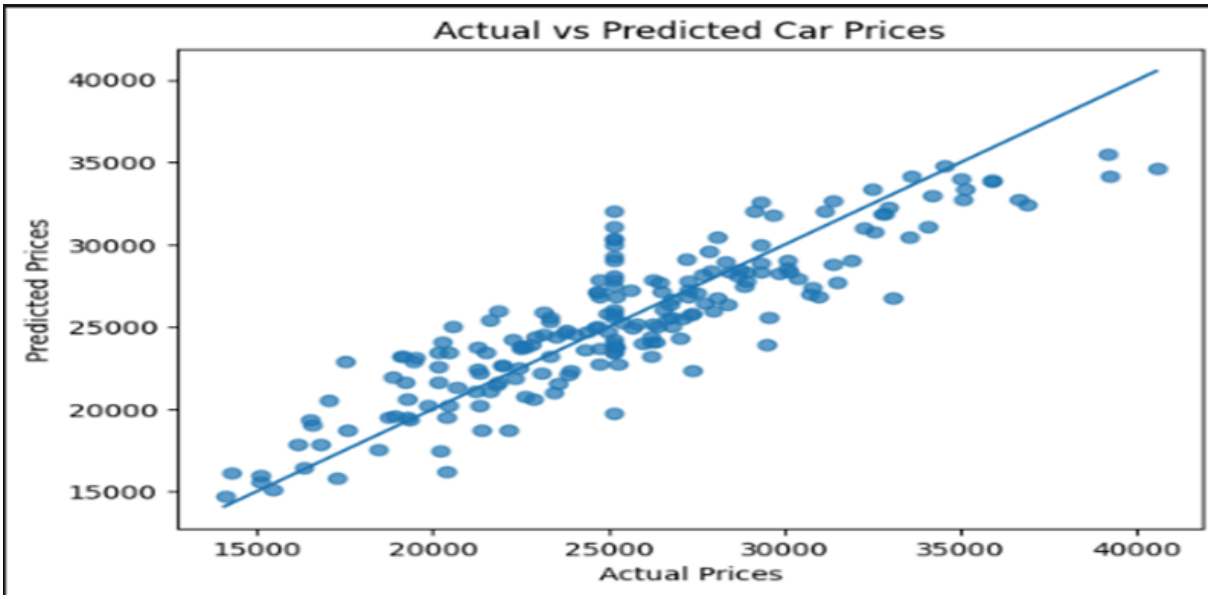
Price by Fuel Type Visualization

The plot shows the median, interquartile range, minimum and maximum values, as well as any outliers, allowing for an effective comparison of price levels and variability across different fuel types.

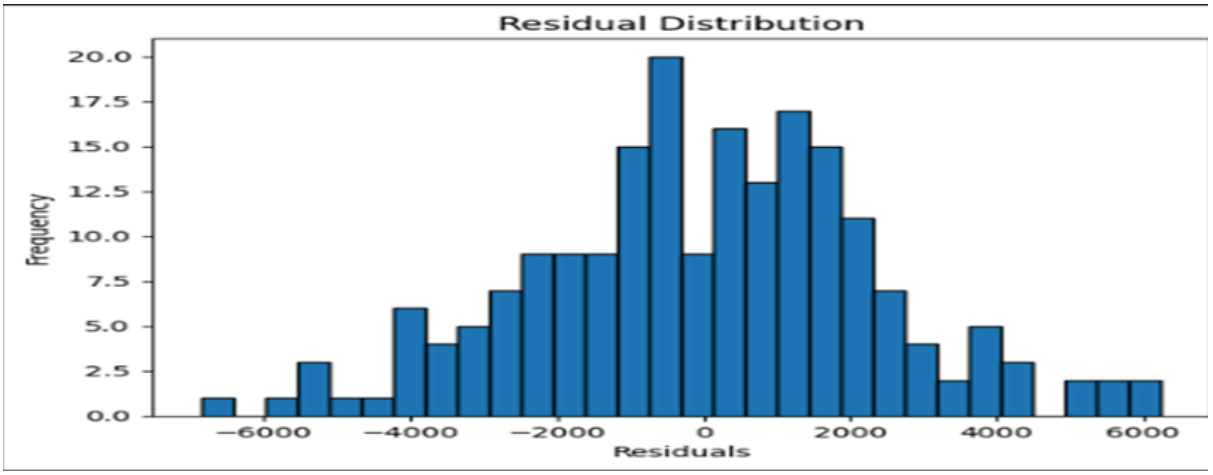


Actual vs Predicted Car Prices Visualization

Scatter plot illustrates how closely predicted prices align with actual values.



Residual Distribution Visualization
Residual histogram evaluates error distribution and checks if residuals are centered around zero.



5. Linear Regression Model

- **Features Used:** Year, Engine Size, Mileage, Make, Fuel Type, Transmission
- **Target:** Price
- **Steps:**
- Split dataset into training (80%) and testing (20%)
- Trained a Linear Regression model
- Predicted prices on test set

6. Evaluation Metrics

- **R^2 Score:** 0.783
- **MSE (Mean Squared Error):** 5,620,317
- **RMSE (Root Mean Squared Error):** $2,370.72

Finally Streamlit User Interface (UI):

A dedicated web application was developed using “Streamlit” to provide a user-friendly and interactive interface for the deployed car price prediction model. Streamlit was chosen for its ability to rapidly create data applications using only Python, simplifying the transition from model development to a functional web tool.

1. Interface Design and Input Fields:

The application presents a clean, single-page interface with input widgets corresponding directly to the features required by the Linear Regression model:

Feature	Streamlit Widget	Purpose
Year	`st.number_input`	Year of manufacture
Engine Size	`st.number_input`	Engine displacement in liters
Mileage	`st.number_input`	Total distance driven in kilometers
Fuel Type	`st.selectbox`	Type of fuel used (Petrol, Diesel, Electric)
Transmission	`st.selectbox`	Gearbox type (Manual, Automatic)
Make	`st.selectbox`	Car manufacturer (Toyota, Ford, Honda, BMW)

Core Prediction Logic

The user initiates the prediction by clicking the "Predict Price" button.

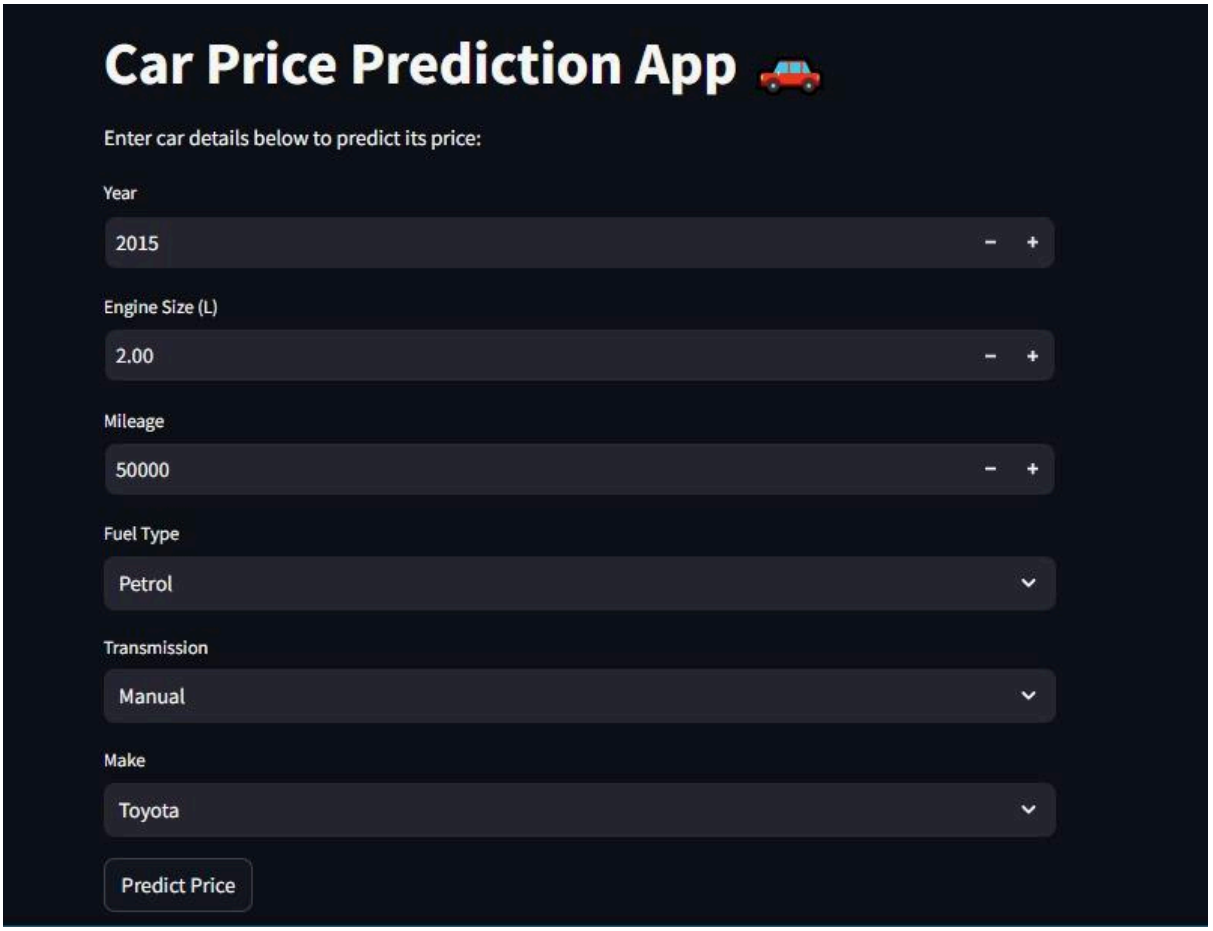
The application then executes the following steps:

1. Data Preparation: The collected inputs are converted into the required DataFrame format, including the one-hot encoding of categorical variables.

2. Data Scaling: The numerical features are scaled using the pre-loaded ‘scaler’.

3.Model Inference: The prepared data is fed into the ‘Linear Regression’ model.

4.Result Display: The predicted price is displayed to the user in a clear, formatted success message, ensuring the price is non-negative.



This interface successfully abstracts the complexity of the machine learning pipeline, allowing any user to easily obtain a market price estimate for a car based on its specifications.